Stochastic Intrinsic Kriging for Simulation Metamodelling

Mehdad, E.; Kleijnen, Jack P.C.

Publication date:
2014

Document Version
Early version, also known as pre-print

CentER

Discussion Paper

No. 2014-054

STOCHASTIC INTRINSIC KRIGING FOR SIMULATION METAMODELLING

By

Ehsan Mehdad, Jack P.C. Kleijnen

8 September, 2014

ISSN 0924-7815
ISSN 2213-9532

TILBURG UNIVERSITY

Stochastic Intrinsic Kriging for Simulation Metamodelling

Ehsan Mehdad and Jack P.C. Kleijnen
Tilburg School of Economics and Management, Tilburg University

September 5, 2014

Abstract

We derive intrinsic Kriging, using Matherons intrinsic random functions which eliminate the trend in classic Kriging. We formulate this intrinsic Kriging as a metamodel in deterministic and random simulation models. For random simulation we derive an experimental design that also specifies the number of replications that varies with the input combinations. We compare intrinsic Kriging and classic Kriging in several numerical experiments with deterministic and random simulations. These experiments suggest that intrinsic Kriging gives more accurate metamodel, in most experiments.

Keywords: Gaussian process, Kriging, intrinsic Kriging, metamodel, computer experiment, simulation

JEL: C0, C1, C9, C15, C44

1 Introduction

Kriging or *Gaussian process* (GP) modeling is popular in geostatistics (Krige himself was a mining engineer), either deterministic or random simulation or computer experiments (our focus), and machine learning. Classic textbooks in these three areas are Cressie (1991), Santner et al. (2003) and Rasmussen and Williams (2006).

Kriging has several variants; we focus on *ordinary Kriging* (OK) and *universal Kriging* (UK). OK assumes that the GP output (dependent variable) Y is the sum of a constant mean β_0 and a second-order stationary GP with zero mean and covariance matrix $\mathbf{\Sigma}_M$. UK replaces this constant mean by some specific function; e.g., a polynomial in the inputs (independent variables). We focus on UK with a polynomial of a fixed order p in the d explanatory variables x_g $(g = 1, \ldots, d)$; e.g., if $p = 2$ and $d = 1$ so $x_1 = x$, then $E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$. Obviously, OK is a special case of UK; i.e., OK is the same as UK with $p = 0$.

If the true input/output (I/O) function of the underlying simulation model $f(\mathbf{x})$ is *monotonic*, then intuitively it seems best to use UK with $p = 1$ or $p = 2$. Unfortunately, in practice we do not know the true values of the $p+1$ regression

coefficients in $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ so we estimate $\boldsymbol{\beta}$ from the simulation's I/O data. This estimate $\hat{\boldsymbol{\beta}}$ creates noise and bias, so in practice we might use OK instead of UK because OK requires the estimation of the scalar β_0 instead of all the elements of the vector $\boldsymbol{\beta}$. Anyhow, in general it is not clear whether OK or UK is best (so we might use cross-validation to estimate the best Kriging metamodel; see the preceding references).

The idea of *intrinsic random functions* (IRFs) is to remove the trend from data by linear filtration of data to remove dependence on the trend (as we shall see in § 3). Originally, IRFs were introduced by the French mathematician George Matheron in his seminal paper Matheron (1973). He not only formalized Kriging but also extended it to IRFs. He developed IRFs to tackle the difficult problem of estimating a variogram in a UK model where a spatial trend creates bias that affects the variogram of the residuals; see Cressie (1991, pp. 299-306) and Chilès and Delfiner (2012, pp. 238-298). It was known before Matheron that in the presence of a polynomial trend of order k, the $(k+1)$th order difference of data annihilate the trend and turn non-stationary data into stationary. Matheron extended this idea for higher-order stationary increments and defined a general class of processes which he called intrinsic random functions of order k (IRF-k). In the time series literature, these IRF-k's are known as integrated processes.

An IRF gives a metamodel with a *generalized covariance matrix* $\mathbf{K}$ which replaces the covariance matrix for the stationary random functions in Kriging. $\mathbf{K}$ needs to be conditionally positive definite and—unlike covariances of stationary processes—$\mathbf{K}$ may be unbounded and can describe processes with unbounded dispersion such as Brownian motion processes.

To the best of our knowledge, IRFs have not yet been applied in simulation, neither deterministic nor random. Therefore it seems interesting to formalize IRFs for a simulation readership, and to compare the performance of Kriging and IRFs in experiments with test functions. For all these experiments we use the root mean square error (RMSE) performance measure. In all these experiments we also study the effects of the number of input combinations. These experiments include simulation of test functions with different dimensionality. The main conclusion is that in both deterministic and random simulation the intrinsic Kriging give a more accurate metamodel than Kriging.

We organize the rest of this paper as follows. Section 2 summarizes UK. Section 3 formalizes IRF-k. Section 4 uses this IRF-k to formalize intrinsic Kriging (IK). Section 5 discusses the selection of the generalized covariance matrix $\mathbf{K}$ in IK. Section 6 adapts IK for random simulation, including a design of experiment for such simulation. Section 7 presents numerical experiments. Section 8 summarizes our conclusions.

2 Universal Kriging

In this section we summarize UK, following Cressie (1991, pp. 151-182). Kriging gives the *optimal* predictor of a random process Y at an unobserved or "new"

point (or "input combination", in simulation terminology) $\mathbf{x}_0$—given the value of Y at the already observed "old" points $\mathbf{x} = (x_1, \ldots, x_d)$. UK assumes

$$Y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + M(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^d, \tag{1}$$

where $\mathbf{f}(\mathbf{x})$ is a vector of $p+1$ regression functions or "trend", $\boldsymbol{\beta}$ is a vector of $p+1$ parameters, and $M(\mathbf{x})$ is a second-order stationary GP with zero mean and covariance matrix $\boldsymbol{\Sigma}_M$.

$\boldsymbol{\Sigma}_M$ must be specified such that it makes $M(\mathbf{x})$ in (1) a second-order stationary GP; i.e., $\boldsymbol{\Sigma}_M$ is a function of the *distance* between the points $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ with $i, i' = 0, 1, \ldots, m$ where the subscript 0 denotes a new point and m denotes the number of old points. *Anisotropic* covariance functions use the distances along the d axes $h_{i;i';g} = |x_{i;g} - x_{i';g}|$ $(g = 1, \ldots, d)$. *Isotropic* functions use the Euclidean distance h between the points in $\mathbb{R}^d$ so $h = \|\mathbf{h}\| = (\sum_{g=1}^d h_g^2)^{1/2}$ (for notational simplicity we suppress the subscripts i and i'); also see (13). Anisotropic functions are popular in simulation. The most popular choice for the covariance function in $M(\mathbf{x})$ is the so-called *Gaussian* covariance function:

$$\text{cov}(\mathbf{x}_i, \mathbf{x}_{i'}) = \tau^2 \prod_{g=1}^{d} \exp\left(-\theta_g h_{i;i';g}^2\right) \text{ with } \theta_g > 0, \tag{2}$$

where τ^2 is the variance of $M(\mathbf{x})$.

Let $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_m))$ denote the m values of the metamodel in (1) at the m old points. Kriging predicts Y *linearly* from the old I/O data $(\mathbf{X}, \mathbf{Y})$ where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is the $d \times m$ matrix with m old input combinations $\mathbf{x}_i = (x_{i;g})$ $(i = 1, \ldots, m;\ g = 1, \ldots, d)$:

$$\hat{Y}(\mathbf{x}_0) = \boldsymbol{\lambda}^\top \mathbf{Y} \text{ such that } \boldsymbol{\lambda}^\top \mathbf{F} = \mathbf{f}(\mathbf{x}_0)^\top, \tag{3}$$

where $\mathbf{F}$ is the $m \times (p+1)$ matrix with element (i, j) being $f_j(\mathbf{x}_i)$, $\mathbf{f}(\mathbf{x}_0) = (f_0(\mathbf{x}_0), \ldots, f_p(\mathbf{x}_0))^\top$, and the condition for $\boldsymbol{\lambda}$ guarantees that $\hat{Y}(\mathbf{x}_0)$ is an *unbiased* predictor. The optimal linear unbiased predictor minimizes the mean squared prediction error (MSPE), defined as

$$\text{MSPE}(\hat{Y}(\mathbf{x}_0)) = E(\hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_0))^2.$$

Cressie (1991, pp. 151-157) shows how to use Lagrangian multipliers to solve this constrained minimization, which gives the *optimal* weights:

$$\boldsymbol{\lambda}^\top = \left(\boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot) + \mathbf{F}\left(\mathbf{F}^\top \boldsymbol{\Sigma}_M^{-1} \mathbf{F}\right)^{-1} \left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \boldsymbol{\Sigma}_M^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)\right)\right)^\top \boldsymbol{\Sigma}_M^{-1} \tag{4}$$

with the m-dimensional vector with covariances between the outputs of the new and the m old points $\boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot) = (\boldsymbol{\Sigma}_M(\mathbf{x}_0, \mathbf{x}_1), \ldots, \boldsymbol{\Sigma}_M(\mathbf{x}_0, \mathbf{x}_m))^\top$, and $\boldsymbol{\Sigma}_M$ the $m \times m$ matrix with the covariances between the outputs of the old points so element (i, i') is $\boldsymbol{\Sigma}_M(\mathbf{x}_i, \mathbf{x}_{i'})$. The resulting minimal MSPE is

$$\begin{aligned} &\text{MSPE}(\hat{Y}(\mathbf{x}_0)) = \tau^2 - \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_M^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot) + \\ &\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \boldsymbol{\Sigma}_M^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)\right)^\top \left(\mathbf{F}^\top \boldsymbol{\Sigma}_M \mathbf{F}\right)^{-1} \left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \boldsymbol{\Sigma}_M^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)\right). \end{aligned} \tag{5}$$

Because the predictor is unbiased, this MSPE equals the predictor variance, which is often called the "Kriging variance".

Note: Kriging is an exact interpolator; i.e., (4) for the old points gives a predictor that equals the observed output. For the old points the Kriging variance (5) reduces to zero.

3 Intrinsic random functions of order k

In this section we formalize the IRF-k, following Cressie (1991, pp. 299-306) and Chilès and Delfiner (2012, pp. 252-257). We rewrite (1) as

$$\boldsymbol{\mathscr{Y}}(\mathbf{x}) = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\mathscr{M}}(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^d, \tag{6}$$

where $\boldsymbol{\mathscr{Y}}(\mathbf{x}) = (\mathscr{Y}(\mathbf{x}_1), \ldots, \mathscr{Y}(\mathbf{x}_m))^\top$, and $\boldsymbol{\mathscr{M}}(\mathbf{x}) = (\mathscr{M}(\mathbf{x}_1), \ldots, \mathscr{M}(\mathbf{x}_m))^\top$. We no longer assume $\mathscr{M}$ is second-order stationary. Let $\mathbf{Q}$ be an $m \times m$ matrix such that $\mathbf{QF} = \mathbf{O}$ where $\mathbf{O}$ is an $m \times (p+1)$ matrix with all elements zero. Together $\mathbf{Q}$ and (6) give

$$\mathbf{Q}\boldsymbol{\mathscr{Y}}(\mathbf{x}) = \mathbf{Q}\boldsymbol{\mathscr{M}}(\mathbf{x}).$$

Consequently, the second-order properties of $\mathbf{Q}\boldsymbol{\mathscr{Y}}(\mathbf{x})$ depend on $\mathbf{Q}\boldsymbol{\mathscr{M}}(\mathbf{x})$ and *not on the regression function* $\mathbf{F}\boldsymbol{\beta}$.

To generalize the model in (1), we need a stochastic process for which $\mathbf{Q}\boldsymbol{\mathscr{M}}(\mathbf{x})$ is second-order stationary; such processes are called *intrinsically* stationary processes. We assume that $f_j(\mathbf{x})$ $(j = 1, \ldots, p+1)$ are mixed monomials $x_1^{i_1} \cdots x_d^{i_d}$ with $\mathbf{x} = (x_1, \ldots, x_d)^\top$ and nonnegative integers $i_1, \ldots, i_d$ such that $i_1 + \ldots + i_d \leq k$ with k a given nonnegative integer. An IRF-k is a random process $\mathscr{Y}$ for which

$$V(\mathbf{x}^*) = \sum_{i=1}^{m} \lambda_i \mathscr{Y}(\mathbf{x}_i + \mathbf{x}^*) \text{ with } \mathbf{x}_i, \mathbf{x}^* \in \mathbb{R}^d$$

is second-order stationary, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top$ is a *generalized-increment* vector of real numbers such that

$$(f_j(\mathbf{x}_1), \ldots, f_j(\mathbf{x}_m))\, \boldsymbol{\lambda} = 0 \ (j = 1, \ldots, p+1).$$

We give the following four examples of IRF-k following Matheron (1973) and Chilès and Delfiner (2012).

1. The $(k+1)$ integral of a zero-mean stationary random function $Z(t)$ is an IRF-k:
$$Z_k(x) = \int_0^x \frac{(x-t)^k}{k!} Z(t)dt.$$

2. Integrating k times a Brownian motion $B(x)$ gives an IRF-k:
$$B_k(x) = \int_0^x \frac{(x-t)^k}{k!} B(t)dt.$$

3. An ARIMA process (autoregressive integrated moving average process) is a process whose finite difference of order k is a stationary ARMA process, so an ARIMA process is an IRF-$(k-1)$.

4. If $Z(x)$ is a random function which is differentiable $(k+1)$ times and if all its derivatives of order $(k+1)$ are stationary with zero mean, then $Z(x)$ is an IRF-k. This example characterizes a differentiable IRF-k; of course there are non-differentiable IRF-k's.

4 Intrinsic Kriging

In this section we introduce *intrinsic Kriging* (IK) based on an IRF-k defined in §3. Let $\mathscr{M}(\mathbf{x})$ be an IRF-k with mean zero and *generalized covariance* matrix $\mathbf{K}$. Then the IK metamodel is:

$$\mathscr{Y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathscr{M}(\mathbf{x}). \tag{7}$$

Cressie (1991, pp. 299-306) derives a linear predictor for the IRF-k metamodel defined in (7) with generalized covariance matrix $\mathbf{K}$ that is the analogue of UK. We have the old outputs $\boldsymbol{\mathscr{Y}} = (\mathscr{Y}(\mathbf{x}_1), \ldots, \mathscr{Y}(\mathbf{x}_m))^\top$. The optimal linear prediction of $\mathscr{Y}$ at a new location $\mathbf{x}_0$ follows from minimizing the MSPE of the linear predictor:

$$\min_{\boldsymbol{\lambda}} E\left(\hat{\mathscr{Y}}(\mathbf{x}_0) - \mathscr{Y}(\mathbf{x}_0)\right)^2 \text{ such that } \hat{\mathscr{Y}}(\mathbf{x}_0) = \boldsymbol{\lambda}^\top \boldsymbol{\mathscr{Y}}. \tag{8}$$

IK should meet the condition $E\left(\hat{\mathscr{Y}}(\mathbf{x}_0)\right) = E\left(\mathscr{Y}(\mathbf{x}_0)\right)$, which is equivalent to

$$\boldsymbol{\lambda}^\top \mathbf{F} = \left(f_0(\mathbf{x}_0), \ldots, f_p(\mathbf{x}_0)\right). \tag{9}$$

Note that this condition is not introduced as the unbiasedness condition but as the condition that guarantees that the coefficients of the prediction error $\lambda_1 \mathscr{Y}(\mathbf{x}_1) + \ldots + \lambda_m \mathscr{Y}(\mathbf{x}_m) - \mathscr{Y}(\mathbf{x}_0)$ create a generalized-increment vector $\boldsymbol{\lambda}_{m+1}^\top = \left(\boldsymbol{\lambda}^\top, \lambda_0\right)$ with $\lambda_0 = -1$. This gives the variance of the IK, denoted by σ^2_{IK}:

$$\sigma^2_{\text{IK}} = \text{var}(\boldsymbol{\lambda}_{m+1}^\top \boldsymbol{\mathscr{Y}}) = \sum_{i=0}^{m} \sum_{i'=0}^{m} \lambda_i \lambda_{i'} \mathbf{K}(\mathbf{x}_i, \mathbf{x}_{i'}). \tag{10}$$

In this section we assume that $\mathbf{K}$ is known, so the optimal linear predictor is obtained through minimization of (10) subject to (9). This problem resembles the previous UK objective function with $\boldsymbol{\Sigma}_M$ now replaced by $\mathbf{K}$. Hence, the IK predictor is given by (8) with

$$\boldsymbol{\lambda}^\top = \left(\mathbf{K}(\mathbf{x}_0, \cdot) + \mathbf{F}\left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1}\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}(\mathbf{x}_0, \cdot)\right)\right)^\top \mathbf{K}^{-1}, \tag{11}$$

where $\mathbf{K}(\mathbf{x}_0,\cdot)=(\mathbf{K}(\mathbf{x}_0,\mathbf{x}_1),\ldots,\mathbf{K}(\mathbf{x}_0,\mathbf{x}_m))^\top$ and $\mathbf{K}$ is an $m\times m$ matrix with (i,i') element $\mathbf{K}(\mathbf{x}_i,\mathbf{x}_{i'})$. The resulting σ^2_{IK} is

$$\text{MSPE}(\hat{\mathscr{Y}}(\mathbf{x}_0))=\mathbf{K}(\mathbf{x}_0,\mathbf{x}_0)-\mathbf{K}(\mathbf{x}_0,\cdot)^\top\mathbf{K}^{-1}\mathbf{K}(\mathbf{x}_0,\cdot)+$$
$$\left(\mathbf{f}(\mathbf{x}_0)-\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{K}(\mathbf{x}_0,\cdot)\right)^\top\left(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\left(\mathbf{f}(\mathbf{x}_0)-\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{K}(\mathbf{x}_0,\cdot)\right). \quad (12)$$

Like UK, IK is an exact interpolator. Comparing the predictor and MSPE of UK and IK shows that if $\mathscr{M}(\mathbf{x})$ is a second-order stationary process, UK and IK give identical results.

5 Choosing a generalized covariance matrix K

In this section we discuss properties of generalized covariance matrices and different functions for them. Following Matheron (1973), Cressie (1991, pp. 304-305), and Chilès and Delfiner (2012, pp. 257-269); we begin with isotropic covariance functions for $\mathbf{K}(h=\|\mathbf{h}\|)$.

Obviously $\mathbf{K}$ is symmetric; i.e., $\mathbf{K}(\mathbf{x}_i,\mathbf{x}_{i'})=\mathbf{K}(\mathbf{x}_{i'},\mathbf{x}_i)$, and it must be *conditionally* positive definite so

$$\text{var}(\boldsymbol{\lambda}^\top\boldsymbol{\mathscr{Y}})=\sum_{i=1}^m\sum_{i'=1}^m\lambda_i\lambda_{i'}K(\mathbf{x}_i-\mathbf{x}_{i'})\geq 0 \text{ such that } (f_j(\mathbf{x}_1),\ldots,f_j(\mathbf{x}_m))\,\boldsymbol{\lambda}=0,$$

where the condition must hold for $j=1,\ldots,p+1$. Parametric models for $\mathbf{K}$ are given by Matheron (1973).

We start with *isotropic polynomial* functions developed by Matheron (1973). $\mathbf{K}$ in an IRF-k with isotropic polynomial covariance functions equals

$$K(h)=\begin{cases}-\theta_1 h, & k=0\\ -\theta_1 h+\theta_2 h^3, & k=1\\ -\theta_1 h+\theta_2 h^3-\theta_3 h^5, & k=2,\end{cases} \quad (13)$$

where for $k=0$, 1, or 2 we have the constraints $\theta_1\geq 0$, $\theta_3\geq 0$, and $\theta_2\geq[(20/3)(1+(2/(d+1))\theta_1\theta_3]^{1/2}$ in $\mathbb{R}^d$; obviously, for $k=0$ we have $\theta_2=0$ and $\theta_3=0$ and for $k=1$ we have $\theta_3=0$. For general k the isotropic polynomial generalized covariance function is

$$K(h)=\sum_{l=0}^k(-1)^{l+1}\theta_{l+1}h^{2l+1} \text{ with } h=\|\mathbf{h}\|\geq 0,$$

where $\theta_1,\ldots,\theta_{l+1}$ must satisfy

$$\sum_{l=0}^k\frac{\theta_{l+1}\Gamma\left((2l+1+d)/2\right)}{\pi^{2l+2+(d/2)}\Gamma\left(1+(1/2)(2l+1)\right)}\rho^{-d-2l+1}\geq 0 \text{ for any } \rho\geq 0,$$

where $\Gamma(\cdot)$ denotes the Gamma function.

We use the first two examples of IRF-k that we gave in § 3, to introduce new generalized covariance functions.

1. The $(k+1)$ integral of a zero-mean stationary random function with covariance $C(h = x - x')$ is an IRF-k with generalized covariance function

$$K(h) = (-1)^{k+1} \int_0^h \frac{(h-u)^{2k+1}}{(2k+1)!} C(u) du.$$

There are many choices for C. The common choices are exponential $C(h) = \exp(-\theta h)$, Gaussian $C(h) = \exp(-\theta h^2)$, or Matern. Now we derive the generalized covariance functions for $k = 0$ in two cases of $C(\cdot)$.

(a) Gaussian $C(\cdot)$:

$$\begin{aligned} K(h) &= (-1) \int_0^h (h-u) \exp(-\theta u^2) du \\ &= \left(1 - \exp(-\theta h^2) - h\sqrt{\pi}\mathrm{erf}(\sqrt{\theta} h)\right) / 2\theta, \end{aligned}$$

where $\mathrm{erf}(\cdot)$ is the error function.

(b) Exponential $C(\cdot)$:

$$\begin{aligned} K(h) &= (-1) \int_0^h (h-u) \exp(-\theta u) du \\ &= \frac{h}{\theta} \left(\exp(-\theta h) - 1\right) + \frac{1}{\theta^2} \left(1 - (1 + \theta h) \exp(-\theta h)\right). \end{aligned}$$

2. Integrating k times a Brownian motion $B(x)$ with covariance function $C(x, x'; \theta) = \theta \min(x, x')$ with $x, x' \in [0, 1]$ and $\theta \geq 0$ gives an IRF-k with generalized covariance function

$$K(x, x'; \theta) = \theta \int_0^1 \frac{(x-u)_+^k (x'-u)_+^k}{(k!)^2} du, \tag{14}$$

where $\theta \geq 0$; see Berlinet and Thomas-Agnan (2004, p. 92). Salemi et al. (2013, pp. 546-547) suggest to add polynomial terms to this covariance to avoid $B(x)$ becoming zero at $x = 0$. In our experiments we add a constant term $\theta_0 \geq 0$ to 14. Note that Brownian motion is a special case of fractional Brownian motion B_H which is an IRF-0 with covariance function

$$\mathrm{cov}\left(B_H(x), B_H(x')\right) = \frac{1}{2}\left(x^{2H} + x'^{2H} - (x - x')^{2H}\right), \qquad 0 < H < 1,$$

where $H = 1/2$ gives the usual Brownian motion.

We are also interested in *anisotropic* generalized functions. We use the same idea which we used to handle anisotropy for covariances of stationary random function; see (2) for the Gaussian case. The idea is that the multiplication of valid covariance functions gives a valid covariance function, so we multiply the covariance functions per input dimension.

Below we give the anisotropic version of K for the covariance function of the integrated Brownian motion defined in (14):

$$K(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \prod_{g=1}^{d} \left(\theta_{0;g} + \theta_{1;g} \int_0^1 \frac{(x_g - u_g)_+^{k_g} (x'_g - u_g)_+^{k_g}}{(k_g!)^2} du_g \right), \tag{15}$$

where $\boldsymbol{\theta} = (\theta_{0;1}, \theta_{1;1}, \theta_{0;2}, \ldots, \theta_{0;d}, \theta_{1;d}) \geq 0$.

The anisotropic covariance function accepts different k for different input dimensions, so we have a vector of the orders $\mathbf{k} = (k_1, \ldots, k_d)^\top$ instead of a single scalar k for all the input dimensions in the isotropic covariance functions. Anisotropic covariance functions handle each input dimension separately and this make them more flexible than isotropic covariance functions. However, this comes with the cost of estimating more parameters.

In practice, $\mathbf{K}$ is unknown so we estimate the covariance function parameters $\boldsymbol{\theta}$. For this estimation we use *restricted maximum likelihood* (REML). So we assume $\mathscr{Y}$ is a *Gaussian* IRF-k. The REML estimator of $\boldsymbol{\theta}$ is then found through minimization of the negative log-likelihood function

$$\begin{aligned} \ell(\boldsymbol{\theta}) = (m-q)/2 \log(2\pi) - \frac{1}{2} \log \left| \mathbf{F}^\top \mathbf{F} \right| + \frac{1}{2} \log |\mathbf{K}(\boldsymbol{\theta})| + \frac{1}{2} \log \left| \mathbf{F}^\top \mathbf{K}(\boldsymbol{\theta})^{-1} \mathbf{F} \right| \\ + \frac{1}{2} \boldsymbol{\mathscr{Y}}^\top \boldsymbol{\Xi}(\boldsymbol{\theta}) \boldsymbol{\mathscr{Y}}, \end{aligned} \tag{16}$$

where $q = \operatorname{rank}(\mathbf{F})$ and $\boldsymbol{\Xi}(\boldsymbol{\theta}) = \mathbf{K}(\boldsymbol{\theta})^{-1} - \mathbf{K}(\boldsymbol{\theta})^{-1} \mathbf{F} \left(\mathbf{F}^\top \mathbf{K}(\boldsymbol{\theta})^{-1} \mathbf{F} \right)^{-1} \mathbf{F}^\top \mathbf{K}(\boldsymbol{\theta})^{-1}$. Finally, we replace $\mathbf{K}$ by $\mathbf{K}(\hat{\boldsymbol{\theta}})$ in (11) to obtain $\widehat{\boldsymbol{\lambda}}$ and in (12) to obtain $\hat{\sigma}^2_{\mathrm{IK}}$.

We could require REML to estimate the optimal (integer) orders $\mathbf{k}^*$, but this would make the optimization difficult. In our methodology, the user has to try different values for $\mathbf{k}$ and pick the one which gives a better fit. The development of a procedure to find $\mathbf{k}^*$ without user intervention is a topic for future research.

6 Stochastic simulation and IK: SIK

In this section we extend the theory of IK to account for a situation where the simulation output is random and its variance changes across the input space. We mentioned earlier that IK is an interpolator; this is not a good property for random simulation. Random simulation has sampling variability or internal noise besides the external noise that is the spatial uncertainty created by the fitted metamodel.

The extension of IK that assumes internal noise with a constant variance of simulation output has already been studied in the literature as nugget effect (geostatistics) or jitter (machine learning). Indeed, Cressie (1991, p. 305) briefly discusses IK in case of a nugget effect, replacing $\mathbf{K}$ by $\mathbf{K} + c_0 \delta(\mathbf{h})$ where $c_0 \geq 0$, $\delta(\mathbf{h}) = 0$ if $\mathbf{h} > 0$, and $\delta(\mathbf{h}) = 1$ if $\mathbf{h} = 0$. Our contribution considers the case of heteroscedastic variance.

Our methodology is similar to the one that is used to incorporate internal noise in Kriging and is published under different names; see Opsomer et al. (1999); Ankenman et al. (2010); Yin et al. (2011).

We suggest a new metamodel which extends the IK metamodel defined in (7) and incorporates the internal noise. The value of this metamodel at replication r of the random output at $\mathbf{x}$ is

$$\mathsf{Y}_r(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}) + \varepsilon_r(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^d, \tag{17}$$

where $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ denotes the internal noise at input combination $\mathbf{x}$. We assume that the internal noise has a Gaussian distribution with mean zero and variance $\mathsf{V}(\mathbf{x})$ and that this internal noise is independent of $\mathsf{M}(\mathbf{x})$.

Our experimental design consists of pairs $(\mathbf{x}_i, n_i),\ i = 1, \ldots, m$, where n_i is the number of replications at input combinations $\mathbf{x}_i$. The replications enable us to compute the classic unbiased estimators of the mean output and the internal variance:

$$\overline{\mathsf{Y}}(\mathbf{x}_i) = \frac{\sum_{r=1}^{n_i} \mathsf{Y}_{i;r}}{n_i} \text{ and } s^2(\mathbf{x}_i) = \frac{\sum_{r=1}^{n_i} \left(\mathsf{Y}_{i;r} - \overline{\mathsf{Y}}(\mathbf{x}_i)\right)^2}{n_i - 1}. \tag{18}$$

Because we assumed that the external noise $\mathsf{M}(\mathsf{x})$ and the internal noise $\varepsilon(\mathbf{x})$ in (17) are independent, the stochastic intrinsic Kriging (SIK) predictor and its MSPE can be derived similarly to the IK in (8) and (12) except that $\mathbf{K}_\mathsf{M}$ will be replaced by $\mathbf{K} = \mathbf{K}_\mathsf{M} + \mathbf{K}_\varepsilon$, where $\mathbf{K}_\varepsilon$ is a diagonal matrix (so no common random numbers are used) with the variances of the internal noise $\mathsf{V}(\mathbf{x}_i)/n_i$ on the main diagonal, and $\mathbf{K}_\mathsf{M}$ still denotes the generalized covariance matrix of IK without internal noise. We also have to replace $\mathscr{Y}$ in (8) and (12) by $\overline{\mathbf{Y}} = (\ \overline{\mathsf{Y}}(\mathbf{x}_1), \ldots, \overline{\mathsf{Y}}(\mathbf{x}_m)\ )^\top$. So the stochastic intrinsic Kriging predictor is

$$\begin{aligned}
&\widehat{\mathsf{Y}}(\mathbf{x}_0) = \boldsymbol{\lambda}^\top \overline{\mathbf{Y}} \text{ where} \\
&\boldsymbol{\lambda}^\top = \left(\mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot) + \mathbf{F}\left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\right)\right)^\top \mathbf{K}^{-1}
\end{aligned} \tag{19}$$

and its MSPE is

$$\begin{aligned}
&\text{MSPE}(\widehat{\mathsf{Y}}(\mathbf{x}_0)) = \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot) + \\
&\left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\right)^\top \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \left(\mathbf{f}(\mathbf{x}_0) - \mathbf{F}^\top \mathbf{K}^{-1} \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\right).
\end{aligned} \tag{20}$$

Note: In our experiments we estimate the MSPE from the known mean output of our test functions; see (23).

We use REML (see § 5) to estimate the parameters of the generalized covariance $\boldsymbol{\theta}$, and replace $\mathbf{K}_\mathsf{M}$ by $\mathbf{K}_\mathsf{M}(\widehat{\boldsymbol{\theta}})$. We also need to estimate the internal noise V which is typically unknown. Inspired by Ankenman et al. (2010) we use an IK metamodel for the internal noise.

$$\mathsf{V}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\sigma} + \mathsf{Z}(\mathbf{x}),$$

where Z is an IRF-k independent of M. We know that $\mathsf{V}(\mathbf{x})$ is not observable, even at old points $\mathbf{x}_i$ $(i = 1, \ldots, m)$, so we let $s^2(\mathbf{x}_i)$ defined in (18) stand for $\mathsf{V}(\mathbf{x}_i)$. We use IK to model the internal noise, so we assume the $s^2(\mathbf{x}_i)$ have no noise and $\widehat{\mathsf{V}}(\mathbf{x}_i) = s^2(\mathbf{x}_i)$. We replace $\mathbf{K}_\varepsilon$ by $\widehat{\mathbf{K}}_\varepsilon = \left(\widehat{\mathsf{V}}(\mathbf{x}_1)/n_1, \ldots, \widehat{\mathsf{V}}(\mathbf{x}_m)/n_m\right)$. Finally, we replace $\mathbf{K} = \mathbf{K}_\mathsf{M} + \mathbf{K}_\varepsilon$ by $\widehat{\mathbf{K}} = \mathbf{K}_\mathsf{M}(\widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{K}}_\varepsilon$ in (19) and (20). In the next section we explain how we choose the number of replications at each point n_i.

6.1 Design of experiments

Like Ankenman et al. (2010) we are interested in an experimental design which has low integrated MSPE (IMSPE) introduced in Sacks et al. (1989), but our approach is slightly different. Ankenman et al. (2010) use MSPE for the case of simple Kriging where the trend is known, whereas we use MSPE for the general case of universal Kriging where the trend is unknown; we also correct an error in Ankenman et al. (2010), as we shall see below (22).

In our design we have to allocate N replications among m old points $\mathbf{x}_i$ such that this design minimizes the IMSPE. Let $\mathscr{X}$ be the design space. Then our goal is

$$\min_{\mathbf{n}} \mathrm{IMSPE}(\mathbf{n}) = \int_{\mathbf{x}_0 \in \mathscr{X}} \mathrm{MSPE}(\mathbf{x}_0, \mathbf{n}) d\mathbf{x}_0, \tag{21}$$

subject to $\mathbf{n}^\top \mathbf{1}_m \leq N$, and $\mathbf{n} = (n_1, \ldots, n_m)^\top$ where $n_i \in \mathbb{N}$.

We formulate IMSPE as

$$\begin{aligned}
&\mathrm{IMSE}(\mathbf{n}) = \int \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \mathbf{x}_0) d\mathbf{x}_0 - \\
&\mathrm{trace}\left\{ \begin{bmatrix} \mathbf{O} & \mathbf{F}^\top \\ \mathbf{F} & \mathbf{K}(\mathbf{n}) \end{bmatrix}^{-1} \int \begin{bmatrix} \mathbf{f}(\mathbf{x}_0)\mathbf{f}(\mathbf{x}_0)^\top & \mathbf{f}(\mathbf{x}_0)\mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \\ \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\mathbf{f}(\mathbf{x}_0)^\top & \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)\mathbf{K}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \end{bmatrix} d\mathbf{x}_0 \right\} \\
&= \int \mathbf{K}_\mathsf{M}(\mathbf{x}_0, \mathbf{x}_0) d\mathbf{x}_0 - \mathrm{trace}\left\{ S(\mathbf{n})^{-1} \int \boldsymbol{\Gamma}(\mathbf{x_0}) d\mathbf{x}_0 \right\} \\
&= \kappa - \mathrm{trace}\left\{ \mathbf{S}(\mathbf{n})^{-1} \mathbf{W} \right\} = \kappa - \sum_{i,i'=1}^{p+1+m} [\mathbf{S}(\mathbf{n})^{-1}]_{i,i'} W_{i,i'} \\
&= \kappa - \mathbf{1}^\top \left[\mathbf{S}(\mathbf{n})^{-1} \circ \mathbf{W} \right] \mathbf{1},
\end{aligned}$$

where $\circ$ is the Hadamard product.

Then we write the Lagrangian function for the minimization problem

$$L(\mathbf{n}) = \mathrm{IMSPE}(\mathbf{n}) + \eta(N - \mathbf{1}^\top \mathbf{n}).$$

The first-order optimality conditions are

$$\frac{\partial L(\mathbf{n})}{\partial n_i} = \frac{\partial \mathrm{IMSPE}(\mathbf{n})}{\partial n_i} + \eta = 0, \qquad (i = 1, \ldots, m).$$

Application of linear algebra gives

$$\frac{\partial \text{IMSPE}(\mathbf{n})}{\partial n_i} = -\mathbf{1}^\top \left[\mathbf{W} \circ \frac{\partial \mathbf{S}(\mathbf{n})^{-1}}{\partial n_i} \right] \mathbf{1}$$
$$= -\frac{\mathsf{V}(\mathbf{x}_i)}{n_i^2} \mathbf{1}^\top \left[\mathbf{W} \circ \left(\mathbf{S}(\mathbf{n})^{-1} \mathbf{J}^{(ii)} \mathbf{S}(\mathbf{n})^{-1} \right) \right] \mathbf{1},$$

where $\mathbf{J}^{(ii)}$ is a $(p+1+m)\times(p+1+m)$ matrix with 1 in position $(p+1+i, p+1+i)$ and zeros elsewhere.

Suppose N is large enough that $\mathsf{K} \approx \mathsf{K}_\mathsf{M}$. We solve the first-order optimality conditions to find $\mathbf{n}^*$. Relaxing the integrality condition, we get the optimal allocation of total number of replications N over m old points:

$$n_i^* \approx N \frac{\sqrt{\mathsf{V}(\mathbf{x}_i) C_i}}{\sum_{i=1}^m \sqrt{\mathsf{V}(\mathbf{x}_i) C_i}}, \text{ with } C_i = \mathbf{1}^\top \left[\mathbf{W} \circ \left(\mathbf{S}^{-1} \mathbf{J}^{(ii)} \mathbf{S}^{-1} \right) \right] \mathbf{1}, \tag{22}$$

so both the internal and the external noises affect the allocation. Note that Ankenman et al. (2010) wrongly simplify further C_i to $\left[\mathbf{S}^{-1}\mathbf{W}\mathbf{S}^{-1}\right]_{qq}$ where $q = p + 1 + m$.

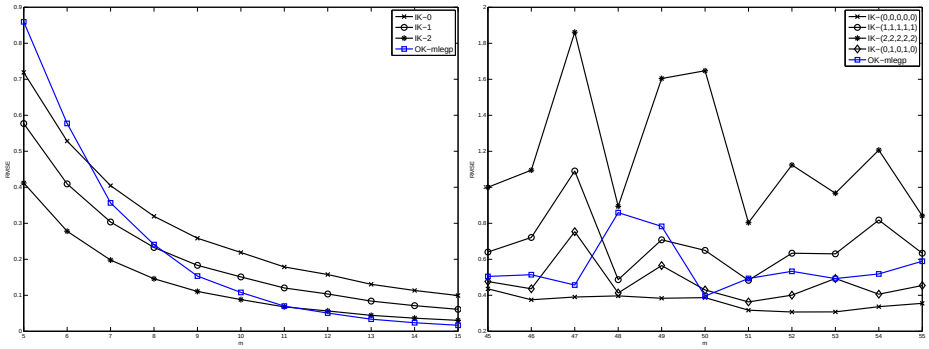
Because we need to estimate K_M and K_ε, we use the *two-stage approach* proposed in Ankenman et al. (2010, p. 378). In Stage 1, we obtain a pilot sample of $m_1 < m$ input combinations and allocate $n_0 > 10$ replications to each point. This enable us to estimate $\mathsf{K}_\mathsf{M}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\mathsf{V}}$. In Stage 2, we first select $m - m_1$ input combinations jointly, and then we optimally allocate the $N - m_1 n_0$ additional replications over m input combinations using (22). In the next section we discuss the application of this approach for the M/M/1 example.

7 Numerical experiments

In this section we present our numerical experiments for both deterministic and random simulations. In all of our experiments we use a zero degree polynomial for the trend ($p = 0$), so UK becomes OK. In deterministic simulation we study the performance of OK versus IK. In random simulation we study the random metamodels developed based on OK and IK which account for internal noise; we call them SK (stochastic Kriging) and SIK.

We tried to use the MATLAB code developed by Ankenman et al. (2010) and Yin et al. (2011) to experiment with these Kriging variants (OK for deterministic simulation and SK for random simulation), but their MATLAB code crashed in experiments with $d > 1$. So we use the R package `mlegp` to implement OK and SK; see Dancik (2013) for more details. We implemented our code for IK and SIK in MATLAB.

Initially we tried several generalized covariance matrices that we discussed in § 5. We decided to use the integrated Brownian motion generalized covariance in (15) for IK and SIK. We use the Gaussian covariance function defined in (2) for OK and SK.



(a) Monotonic one-dimensional function (b) Ackley-5 function

Figure 1: RMSE versus m with IK-$\mathbf{k}$ and OK

We study several values for the number of old points m; namely, all integers between $10d-5$ and $10d+5$, where $m = 10d$ is popular in Kriging; see Loeppky et al. (2009).

We experiment with several values for d (input dimensionality). The simplest case is $d = 1$, so this case may give valuable intuitive insight (obviously, for $d = 1$ isotropic and anisotropic covariance functions are the same). In practice, however, $d > 1$ so we also study test functions with such d values.

To evaluate the performance of OK versus IK and SK versus SIK, we select $m_0 = 100d$ new points $\mathbf{x}_0$. For $d = 1$ we select m and m_0 equispaced points; for $d > 1$ we use Latin hypercube sampling (LHS) to select m and m_0 space-filling points. For this LHS we use the MATLAB function `lhsdesign`.

We quantify the performance of different metamodels through RMSE for the m_0 new points:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{m_0}(Y_t - \widehat{Y}_t)^2}{m_0}}. \tag{23}$$

7.1 Deterministic simulation experiments

In this (sub)section we present the results of our experiments with deterministic test functions of different dimensionality; namely $d = 1, 2, 3, 5$.

We start with a *monotonic* function with $d = 1$; namely, $f(x) = 1/[x(x-1)]$ and $1 < x \leq 2$, which inspired by the mean steady-state waiting-time as a function of the service rate x with the arrival rate equal to 1 in a single-server queue with Markovian arrival and service processes (denoted by M/M/1). This function increases monotonically, and increases drastically as $x \downarrow 1$.

For this function, Figure 1a shows the RMSE of IK-k and OK versus different values for m. We observe that IK with $k = 2$ performs better than OK for $m \leq 11$; for larger m the difference between IK-2 and OK becomes small. Next we experiment with several functions that are popular in optimization; see Dixon and Szego (1978) and `http://www.sfu.ca/~ssurjano/index.html`.

We experiment with (1) Six-hump camel-back with $d = 2$ (2) Hartmann-3 with $d = 3$ (3) Levy-3 with $d = 3$ (4) Ackley-5 with $d = 5$. We define the first three functions in the appendix, and discuss the Ackley-5 function in detail. The Ackley-5 function is defined as

$$f(\mathbf{x}) = -20 \exp\left(-0.2\sqrt{\frac{1}{5}\sum_{i=1}^{5} x_i^2}\right) - \exp\left(\frac{1}{5}\sum_{i=1}^{5}\cos(2\pi x_i)\right) + 20 + \exp(1),$$

with $-2 \leq x_i \leq 2,\ i = 1, \ldots, 5$.

For this function, Figure 1b shows that IK-$(0, 0, 0, 0, 0)$ gives the lowest RMSE for all values of m. We also experimented with different $\mathbf{k}$, and found that the RMSE deteriorated compared with $\mathbf{k} = (0, 0, 0, 0, 0)^\top$.

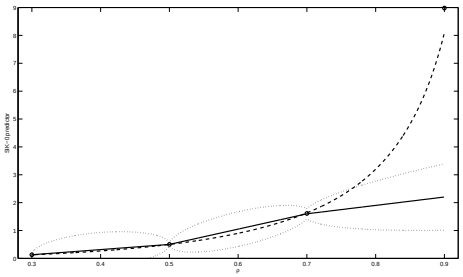
The results for the other test functions show that IK performs better than OK except for the Hartmann-3 function; to be the winner, IK needs an appropriate value for the parameter $\mathbf{k}$ (which is not always $\mathbf{k} = \mathbf{0}$). The figures for these functions can be found in Figure 4 in the appendix.

7.2 Random simulation experiments

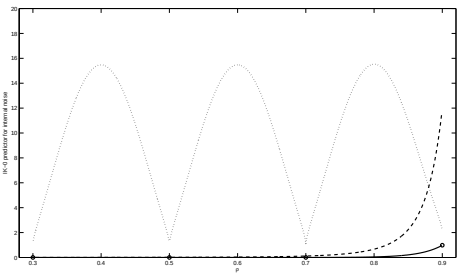
In this (sub)section we first discuss the two-stage approach detailed in § 6.1 for the M/M/1 model (defined in § 7.1) to show how to use SIK in real applications. Then we compare the performance of SIK and SK for different test function simulations. However, in the comparison of SIK and SK for these functions, we do not use the two-stage approach to compute the optimal number of replications at each point; instead we select n_i proportional to the true value of the internal noise at each point. The reason for this approach is that we focus on comparing SIK and SK, and not on the selection of the number of replications.

We design our M/M/1 experiment as follows. We fix the arrival rate at 1 and vary the service rate. This gives the traffic rate $\rho = 1/x$. In each replication we calculate the average waiting time of customers in the system from time 0 to $T = 3,000$ (run-length). To avoid the tactical problem of selecting an initial state and a warm-up period, we do not start in the empty state but we start in the steady state; i.e., the initial number of customers in the system at $T = 0$ equals the mean steady-state number $\rho/(1 - \rho)$. We keep the run-length per replication T the same for all x, so the internal noise is controlled through the number of replications.

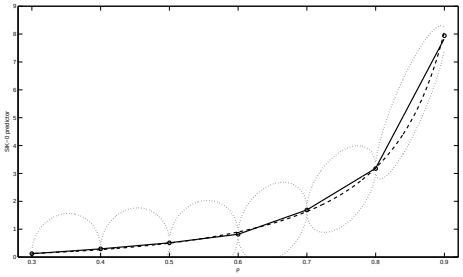
In Stage 1, we select $\rho = 0.3, 0.5, 0.7, 0.9$ and make 20 replications of length T at each point. Then we fit a SIK-0 metamodel to the simulated data. Figure 2a shows the the average waiting time in the queue metamodel from Stage 1. The circles represent the simulated data, the solid curve represents the SIK-0 metamodel, surrounded by $\pm\sqrt{\text{MSPE}(\widehat{\text{Y}}(x_0))}$. The dashed curve represents the true function.



(a) Stage 1



(b) IK-0 metamodel for the internal noise



(c) Stage 2

Figure 2: Two-stage approach for the M/M/1 queue; the dotted curve denotes the true function, and the solid curve denotes the metamodel

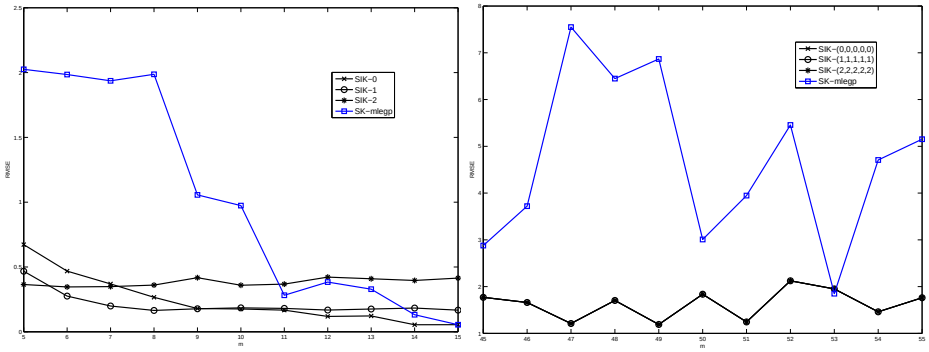
SIK is not an interpolator, so it does not pass through the data points, especially at $x = 0.8$. Note that the $\pm\sqrt{\text{MSPE}(\widehat{\mathsf{Y}}(x_0))}$ intervals which account for internal and external noise cover the true function except for the region between $x = 0.7$ and $x = 0.9$. Figure 2b shows the IK-0 metamodel for the internal noise $\mathsf{V}(x_0)$. Here the metamodel is an interpolator and goes through the data points. It can be seen that the simulation gives a poor estimate of $\mathsf{V}(0.9)$. This figure shows the classic shape of these intervals; namely, their length is close to zero at the old points, and increases away from these points; however, Figure 4 in Ankenman et al. (2010) shows "nearly constant" length.

In Stage 2, we use the information obtained in Stage 1 to apply (22) and optimally allocate $N = 500$ replications over the four old points and three new points $x = 0.4, 0.6, 0.8$. The metamodel for the internal noise help us for these three new points which were not simulated in Stage 1. The estimated optimal allocation is $n^* = 1, 1, 2, 3, 12, 126, 355$. In some points we have already simulated more replications than is optimal, and in other points we have to simulate additional replications. We fit a SIK-0 metamodel to data obtained from Stage 2. Figure 2c shows the SIK-0 metamodel for the data obtained from Stage 2. This figure shows that the new metamodel is close to the true function, and the MSPE intervals cover the true function.

We continue this section with the comparison of SIK and SK in our test functions. In all the test functions we select the number of replications at

each point proportional to the true value of internal noise. In the M/M/1 example, the internal noise function is $\mathsf{V}(x_i)/T \approx 4/(Tx_i(1-1/x_i)^4)$; see Whitt (1989, p. 1350) so $n_i = B\lfloor \mathsf{V}(x_i)/\sum_{i=1}^{m} \mathsf{V}(x_i) \rfloor$ where B is the total number of replications. In the other test functions with higher dimensionality we augment the deterministic response with heteroscedastic noise; namely, $\mathsf{V}(\mathbf{x}_i) = (1 + |y(\mathbf{x}_i)|)^2$, like Wan et al. (2010) doe in an experiment with linear regression metamodels.

Figure 3a shows the RMSE of SIK-k and SK versus m for the M/M/1 simulation. SIK-0 gives smaller RMSE for all m. Figure 3b shows the results for all three $\mathbf{k}$ values for the Ackley-5 simulation. We also observe that changing k from 0 to 1 or 2 in all coordinates give almost identical results. In other test functions SIK-$\mathbf{0}$ gives smaller RMSE in almost all m points except in the Levy-3 function where SIK and SK give almost identical RMSEs for all m. The figures for these test function can be found in Figure 4 in the appendix.



(a) M/M/1 queue (b) Ackley-5

Figure 3: RMSE versus m with SIK and SK

8 Conclusions

Using IRFs, we derived IK for deterministic simulation, and SIK for random simulation. Next we numerically compared the performance—measured by RMSE—of IK with OK and SIK with SK. Additionally, we derived a two-stage approach for sample size allocation in SIK in random simulation. The main conclusion is that in most experiments IK and SIK give smaller RMSEs than OK and SK.

In general, it is well-known that Kriging requires the selection of a specific covariance function (e.g., Gaussian or exponential) and parameters (e.g., $\boldsymbol{\theta}$). In IK and SIK we also must select a specific covariance matrix $\mathbf{K}$, and parameter value $\mathbf{k}$. Theses issues require more research.

Acknowledgement

We thank Barry L. Nelson (Northwestern University) for his MATLAB code for the M/M/1 queue simulation model, and Peter Salemi (Northwestern University) for sharing his R code for the integrated Brownian covariance functions.

A Test functions with $d > 1$

In this appendix we define the test functions with $d > 1$, and give the results of our simulation experiments.

1. Six-hump camel-back with $-2 \le x_1 \le 2,\ -1 \le x_2 \le 1$

$$f(x_1, x_2) = 4x_1^2 - 2.1x_1^4 + x_1^6/3 + x_1x_2 - 4x_2^2 + 4x_2^4$$

2. Hartmann-3 function with $0 \le x_i \le 1,\ i = 1, 2, 3$

$$f(x_1, x_2, x_3) = -\sum_{i=1}^{4} \alpha_i \exp[-\sum_{j=1}^{3} A_{ij}(x_j - P_{ij})^2],$$

with $\boldsymbol{\alpha} = (1.0, 1.2, 3.0, 3.2)^\top$ and A_{ij} and P_{ij} given in Table 1.

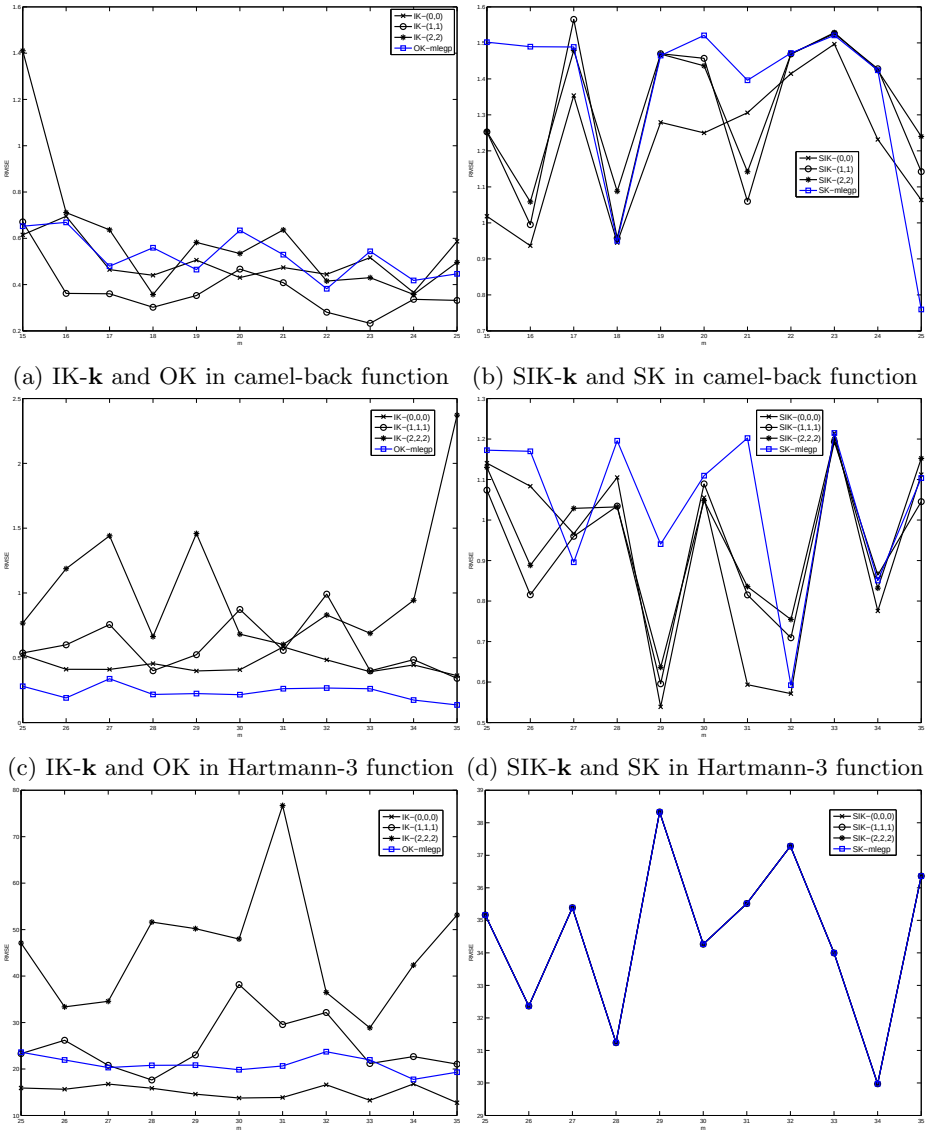
Table 1: Parameters A_{ij} and P_{ij} of the Hartmann-3 function

A_{ij}			P_{ij}		
3	10	30	0.36890	0.1170	0.26730
0.1	10	35	0.46990	0.43870	0.74700
3	10	30	0.10910	0.87320	0.55470
0.1	10	35	0.03815	0.57430	0.88280

3. Levy-3 function with $-10 \le x_i \le 10,\ i = 1, 2, 3$

$$f(\mathbf{x}) = \sin^2(\pi w_1) + (w_2 - 1)^2[1 + 10\sin^2(\pi w_2 + 1)] + (w_3 - 1)^2[1 + \sin^2(2\pi w_3)],$$

where $w_i = 1 + \dfrac{x_i - 1}{4}$.



(a) IK-**k** and OK in camel-back function (b) SIK-**k** and SK in camel-back function

(c) IK-**k** and OK in Hartmann-3 function (d) SIK-**k** and SK in Hartmann-3 function

(e) IK-**k** and OK in Levy-3 function (f) SIK-**k** and SK in Levy-3 function

Figure 4: RMSE versus m in deterministic (left panels) and random simulation (right panels)

References

Ankenman, B. E., Nelson, B. L., and Staum, J. (2010). Stochastic Kriging for simulation metamodeling. *Operations Research*, 58:371–382.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

Chilès, J. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley.

Cressie, N. (1991). *Statistics for spatial data*. Wiley, New York.

Dancik, G. M. (2013). *mlegp: Maximum Likelihood Estimates of Gaussian Processes*. R package version 3.1.4.

Dixon, L. C. W. and Szego, G. P., editors (1978). *Towards Global Optimisation 2*. Elsevier Science Ltd.

Loeppky, J. L., Sacks, J., and Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51:366–376.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468.

Opsomer, J. D., Ruppert, D., Wand, M. P., Holst, U., and Hossjer, O. (1999). Kriging with nonparametric variance function estimation. *Biometrics*, 55(3):704–710.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*. The MIT Press, Cambridge, Massachussets.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Salemi, P., Staum, J., and Nelson, B. (2013). Generalized integrated brownian fields for simulation metamodeling. In *2013 Winter Simulation Conference (WSC)*, pages 543–554.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York.

Wan, H., Ankenman, B. E., and Nelson, B. L. (2010). Improving the efficiency and efficacy of controlled sequential bifurcation for simulation factor screening. *INFORMS Journal on Computing*, 22(3):482–492.

Whitt, W. (1989). Planning queueing simulations. *Management Science*, 35(11):1341–1366.

Yin, J., Ng, S., and Ng, K. (2011). Kriging meta-model with modified nugget effect: an extension to heteroscedastic variance case. *Computers and Industrial Engineering*, 61(3):760–777.